Error bounds for the quadrature using Thiessen's method

Carlos López-Vázquez

LatinGEO Lab IGM+ORT, Universidad ORT Uruguay, Montevideo, Uruguay

carloslopez@uni.ort.edu.uy

ORCID iD: http://orcid.org/0000-0002-8444-1510

SCOPUS AUTHOR ID 55672775500

ResearcherID: http://www.researcherid.com/rid/K-7454-2014

Abstract:

Very frequently in experimental sciences a mean value in a given domain of an empirical function is required. Unlike the traditional quadrature problem, the function is not explicit, and its values are only available at sparsely distributed points thus precluding the application of well established rules. Most methods first interpolate the data points and afterwards apply a numerically exact quadrature over the interpolant. That might explain why they do not provide an estimate of the error committed, as it is customary in numerical analysis. However, the accuracy of the interpolant is not considered. One of such procedures using the nearest neighbour interpolant is due to Thiessen since 1911, and we will show that an error estimate of the quadrature can be built. The procedure was applied to analytical functions in 1D and 2D domains, and the numerical results compared very well with exact values.

MSC: 41A55 Approximate quadratures, 65D32 Quadrature and cubature formulas

Error bounds for the quadrature using Thiessen's method

Very frequently in experimental sciences a mean value in a given domain of an empirical function is required. Unlike the traditional quadrature problem, the function is not explicit, and its values are only available at sparsely distributed points thus precluding the application of well established rules. Most methods first interpolate the data points and afterwards apply a numerically exact quadrature over the interpolant. That might explain why they do not provide an estimate of the error committed, as it is customary in numerical analysis. However, the accuracy of the interpolant is not considered. One of such procedures using the nearest neighbour interpolant is due to Thiessen since 1911, and we will show that an error estimate of the quadrature can be built. The procedure was applied to analytical functions in 1D and 2D domains, and the numerical results compared very well with exact values.

Keywords: quadrature rules, Thiessen polygons, error bounds

Subject classification codes:

1. Introduction

The mean areal value, defined as the ratio of the integral over a domain of the function of interest divided by the area of the domain, is required for many applications in geosciences. However, and unlike other well defined mathematical problems, the function itself is only known at sparse locations thus precluding the use of standard mathematical approaches. More than a century ago, in his seminal paper Thiessen [1] suggested a method to estimate the mean areal value of an experimental variable (in particular, daily precipitation records). With the limited computation tools of the time, the possibilities at hand were just to average the available records or using graphical aids to derive some polygons and use its areas as weights in the linear combination. Mere averaging does not consider at all the closeness of some of the data points, a fact that might attach a large weight to almost duplicated readings. With variants, this

problem is still today very common in the geosciences. It can be characterized by a) data is expensive to acquire, and is available in a limited number of points b) in many cases, location of data points cannot be prescribed in advance c) neither the function itself is explicitly known, nor its partial derivative.

When the function is explicit, or even if it is available through a complex routine or simulation, well established methods are at hand. State of the art routines start evaluating the function at regular intervals in the domain. Afterwards, such values are interpolated with a piecewise, low order interpolant polynomial, whose integral can be obtained exactly. The method is said to be of order p if it is exact for any polynomial function of order p, but not for order p+1. If two methods of different order p and q are available, an estimate of the numerical error can be obtained. If the error is larger than a prescribed tolerance, the domain is further subdivided and new estimates of the integral and its error bound are obtained. The process continues until a stopping criterion is fulfilled: either the bound is below a predefined accepted tolerance, or the prescribed maximum number of function evaluations has been attained [2].

In our setting we have a different situation. There is no freedom to evaluate the function at arbitrary points, neither it is (typically) possible to select their locations. Thiessen [1] offer an estimate but did not provide an error bound for it. To the best of our knowledge, over one century of frequent use of the method did not produce an estimate either. The abovementioned seminal paper describes a method which in present day terms can be described as the exact integral of the nearest neighbour interpolant. In two dimensions (2D), the method relies on a subdivision of the plane into regions, known as Thiessen polygons, Voronoi sets or as Dirichlet tesselation. Its computation has been until today a subject of interest (see, for example, the yearly series of International Symposium on Voronoi Diagrams in Science and Engineering (ISVD)

starting in Tokyo (2004) through St. Petersburg (2013)). Despite its popularity, no known error analysis is available for the Thiessen estimate, a problem which will be addressed here.

This paper is organized as follows. Section 2 reports on the related literature, while Section 3 describes the theory of our proposal. Section 4 discusses its applicability considering some practical aspects. Section 5 describes the numerical experiment and its results, while Section 6 presents the conclusions.

2. Related work

The literature covering error estimates of the quadrature using sparsely located data is surprisingly scarce. For the 1D case Secrest [3] offers analytical bounds for the quadrature assuming that the function has bounded partial derivatives up to degree r. The simplest case is for r=1, but even in that case the bounds should be provided by the user thus limiting its practical applicability. Another possible more general approach is to interpolate with two methods of different accuracy, apply an exact quadrature for both and use their difference as a crude estimate of the error. For example, Gill and Miller [4] used a cubic interpolant, and the error estimate is obtained by subtracting the quadrature estimate using a fourth degree interpolant. The algorithm is still in use; see for example the D01GAF routine documentation from NAG [5]. To the best of our knowledge there is no 2D extension of this method. Singh and Thorpe [6] generalized the Simpson's 1/3 rule and its error estimates for unequally spaced values. Again the bounds for the fourth derivative should be provided by the user. Minaoui *et al.* [7] also considered the case for 1D quadrature. Their rationale assumes that available data points can be considered as displaced from the Gauss-Legendre locations, and afterwards they update the standard quadrature weights. They reported some experiments with radar data.

For the 2D case we found no systematic analysis in the literature. It is a surprise, because it is not difficult to devise some choices for a pair of suitable methods using different interpolants. For example, one might be the nearest neighboor interpolant (which is exact only for constant functions) and the other can be the Triangular Irregular Network with piecewise linear interpolant (which is exact for linear functions). The integration of the first one leads to the Thiessen's estimate, and the other to the 2D Trapezoidal one. Despite its conceptual simplicity we found no reference in the literature of such approach.

Deterministic solutions are not the single possibility for this problem. Provided that the mere average of the experimental values is also an estimate of the unknown areal average, under mild assumptions a valid approach to consider it is the Bootstrap [8]. It can provide not only an estimate, but also a confidence interval for the mean value. Again, despite its conceptual simplicity, we found no example of using it in the case of mean areal estimate. In this case, the overwhelming majority of the literature uses the Thiessen's method, so it is appropriate to derive a direct estimate for it.

3. Theory and Methods

This section illustrates the theory with an example for the unidimensional case (1D), which it is simpler to explain but can be easily extended to higher dimensions. If we subdivide the region of interest (interval [a,b] or domain Ω in general) into N subdomains with extremes s_i we can write that

$$I_{[a,b]} = \int_a^b f(x)\,dx = \sum_{i=1}^{N} \int_{s_{i-1}}^{s_i} f(x)\,dx = \sum_{i=1}^{N} I_{[s_{i-1},s_i]} \tag{1}$$

In the case of the nearest neighbor method, each s_i is chosen as the midpoint of the interval defined by data points $[x_i, x_{i+1}]$. The function f(x) is not explicitly known, but

in this paper it will be assumed that it is so regular that the Taylor expansion up to certain order p+1 is valid. For example, if we set order p=2 we have for each subdomain

$$I_{[s_{i-1},s_i]} = \int_{s_{i-1}}^{s_i} \left[f(x_i) + f'(x_i)(x-x_i) + f''(x_i)\frac{(x-x_i)^2}{2} + f^{(3)}(\xi_i)\frac{(x-x_i)^3}{6} \right] dx = \\ f(x_i)\int_{s_{i-1}}^{s_i} 1dx + f'(x_i)\int_{s_{i-1}}^{s_i} (x-x_i)dx + f''(x_i)\int_{s_{i-1}}^{s_i} \frac{(x-x_i)^2}{2}dx + \int_{s_{i-1}}^{s_i} f^{(3)}(\xi_i)\frac{(x-x_i)^3}{6}dx \tag{2}$$

The first term is what the 1D Thiessen method provides, and it is readily computable. The other terms are products of the higher order derivatives (not known) multiplied by the integral of a monomial which can be explicitly computed. The final expression for each subdomain is

$$I_{[s_{i-1},s_i]} = f(x_i)\Omega_i + \sum_{j=1}^{p} a_{i,j} f^{(j)}(x_i) + R_{[s_{i-1},s_i]} \tag{3}$$

being

$$\Omega_i = \int_{s_{i-1}}^{s_i} dx \ , \ a_{i,j} = \int_{s_{i-1}}^{s_i} \frac{(x-x_i)^j}{j!}dx \text{ and } f^{(j)}(x_i) = \left.\frac{d^{(j)}f(x)}{dx^j}\right|_{x=x_i}$$

In the expression we explicitly included the Thiessen estimate, the contribution of the higher order derivatives, and a residual term. If the higher order derivatives (j=1, 2,.., p) of the empirical function evaluated at data points are available (an unusual but convenient situation to be considered again later), the estimate is easy to build because all terms a_{ij} are computable, and if the Taylor expansion converges we can assume that the residual is small. In the typical situation such higher order derivatives are not available, so they need to be estimated as well. The factors a_{ij} are constant (i.e. independent of function f(x)) but dependent on the integration domain and location of

the data points. The residual $R_{[s_{i-1},s_i]}$ is a multiple of the derivative of order (p+1) evaluated at an unknown location ξ_i belonging to the interval $[s_{i-1}, s_i]$. Summing up, in a typical setting we have 1+N*(p+1) unknowns, including the exact integral I (a scalar), N times a set of higher order derivatives from order 1 to p, and N residual terms. If we denote as R the sum of all residuals, we can write down the following working expression:

$$I = \sum_{i=1}^{N} \left(f(x_i)\Omega_i + \sum_{j=1}^{p} a_{i,j} f^{(j)}(x_i) \right) + R \tag{4}$$

This expression holds valid for any number and location of data points $\{x_i\}$. A crucial observation is that the integral I can be estimated using all the data points, or we can ignore some of them as well. For example, if we leave just one of them out, we can write down up to N equations involving the same set of unknowns. If we leave two of them, we will have N(N-1)/2 equations. If we leave out K points (K<N), there will be $N!/[(N-K)!K!]$ equations, being K! a shorthand for the factorial of K, defined as K!=K(K-1)(K-2)...4.3.2. If N is large enough, removing a few data points might not affect dramatically the order of magnitude of the cummulated residuals $R = \sum_{i=1}^{N} R_{[s_{i-1},s_i]}$ for different choices of K data points, and thus we can assume that they are comparable for a given p. Notice that R is both a function of the order p and on the location of the selected (N-K) data points.

Now we are ready to present the procedure. For convenience we will introduce an unknown vector z holding all the derivatives of function f at data points and the exact integral I. Its transpose will be

$$\left[f^{(1)}(x_1) \quad f^{(2)}(x_1) \quad \ldots \quad f^{(p)}(x_1) \quad f^{(1)}(x_2) \quad f^{(2)}(x_2) \quad \ldots \quad f^{(p)}(x_2) \quad \ldots \quad f^{(p)}(x_N) \quad I \right] \tag{5}$$

If we select a subset of the available data points, the traditional Thiessen estimate and the corresponding factors a_{ij} can be evaluated. The selection can be done in many different ways, and we can consider all the related equations. The linear system can be expressed as $Az + Bf = r$, being A the factor matrix holding the factors a_{ij}; B holds the standard Thiessen weights which properly multiplied by the known function value vector f will produce the traditional Thiessen estimate, and r is a residual vector. If as a first subset we use all data points, first row of matrix A will have the following structure

$$\left[a_{11} \quad a_{12} \quad \ldots \quad a_{1p} \quad a_{21} \quad a_{22} \quad \ldots \quad a_{2p} \quad \ldots \quad a_{Np} \quad -1 \right] \tag{6}$$

and first row of matrix B will have $\left[\Omega_1 \quad \Omega_2 \quad \ldots \quad \Omega_N \right]$. If the second subset ignores at least one data point, some of the elements a_{ij} will change, but some will remain the same as those of the first row. Due to the removal, some zero entries appear in both matrix A and B. For example, if for the second subset we remove the first data point, the second row of A will look as

$\left[0 \quad 0 \quad \ldots \quad 0 \quad a_{21}^{\{2\}} \quad a_{22}^{\{2\}} \quad \ldots \quad a_{2p}^{\{2\}} \quad \ldots \quad a_{Np}^{\{2\}} \quad -1 \right]$ while second row of B will look as

$\left[0 \quad \Omega_2^{\{2\}} \quad \ldots \quad \Omega_N^{\{2\}} \right]$. Notice that we introduced a superscript $\{2\}$ to denote the *second subset*; it is mandatory because, in principle, some row elements involved can be different from those of the first row despite many might be identical.

Since every subset has its corresponding residual we cannot on principle have as many equations as unknowns. To close the system we should add further constraints. All elements of vector r are residuals which we will assume are of the same order of magnitude. Hopefully they will be small if the truncated Taylor expansion is similar to

the unknown function f(x) and p is large enough. In any case we can request the norm of the residual vector r to be minimal when varying z. By selecting a suitable value of K we can specify an overdetermined system of equations, with more equations than unknowns. With the abovementioned structure we can build an ordinary least squares problem, which might provide estimates of the integral I but also of the higher order derivatives at data points. With the calculated value of the "exact" integral I and the available standard Thiessen estimate we can provide both an improved value as well as an error estimate. The rationale behind it is as follows.

The standard Thiessen estimate is exact just for constant functions (p=0). For functions which are themselves polynomials of order q<p, the residual term should be exactly zero, so on principle we can state that our procedure *is of order 1+p*, borrowing from traditional numerical analysis the concept. If our estimate is of a substantially larger order, and if the length scale of the Thiessen regions goes to zero, we can say that in the limit the absolute difference between the low order and high order estimate can be used as an error bound of the high order estimate. Such assumption is commonly used in numerical analysis. A word of caution should be raised, because here we are not in the same situation: the abovementioned length scale will not go to zero because the number N cannot be arbitrarily large. We will show, however, that the rule works fairly well in the simulations even for low to moderate N.

4. Discussion of the numerical procedure

We are interesting in practical application of this approach. Building and solving this overdetermined system of equations is not trivial, so we want to discuss now some of the issues found as well as its possible solutions.

1. Practical issues for repeated use of the procedure

As considered by early applications of the Thiessen approach (like in meteorology) the

location of data points is fixed and known in advance. Thus, the Thiessen weights can be calculated once for a given region, and used many times without recalculations. In our approach and under the same assumptions the matrices A and B can also be calculated once and stored for later use with different data values f. Their size is modest: matrix A has N*P(p) columns and an undetermined number of rows, being P(p) the number of partial derivatives of a function up to order p. Matrix B is sparse, and will have a few non-zero entries. The computing time is spent mostly in building matrix A; given vector f, the solution of the system is rather fast, and certainly amenable for routine use.

2. Handling of missing values

For traditional Thiessen use, when some data point is not available the weights need to be recalculated. Despite its complexity, our approach is more flexible. Unlike traditional Thiessen weights, given matrix A and B our procedure to consider missing elements in vector f is simple and fast. Assume that for a particular date the j-th data point (for example, a weather station) is unavailable. We observe that some rows of A were already built assuming that the j-th data point is not included. They are denoted by having zeros in certain columns, a fact clearly noticeable without resorting to the original list of points. Thus, if one data point is missing, we can easily extract a new matrix A' from A, selecting those rows which has certain columns with exactly zero value. A new matrix B' can also be extracted by collecting the corresponding rows from B. Given A' and B' the numerical problem is almost the same as before. On principle we can handle with this procedure up to K missing values; however, we must acknowledge that if there are an important number of missing values it might be worthwhile to expand the A matrix by increasing K in order to provide a meaningful solution.

3. Ill-posedness of the system of equations

We realized quickly that the system of equation of this problem is ill posed. Such systems of equations are characterized by having solutions that vary wildly even with very tiny changes of elements of either matrix A or right hand side b. The norm of the solution vector can be very large, which in many cases results in useless estimates. Solving problems like these require special methods of regularization (see for example, [9]). All of them find the numerical solutions of a nearby problem, as close (in some sense) to the original one as possible but with additional constraints. For example, one very popular approach is known as Tikhonov regularization. The original problem can be presented, without loss of generality, as a minimization one:

$$\min_z \|b - Az\|_2^2 \tag{7}$$

If, as an additional constraint, we require that the solution norm be bounded, we can solve the nearby problem

$$\min_z \left(\|b - Az\|_2^2 + \lambda^2 \|z\|_2^2 \right) \tag{8}$$

where the parameter λ is a small constant to be estimated. The limit case λ=0 recovers the original problem. Implicitly, this approach assumes that the solution vector z has elements of the same order of magnitude but in our setting this is not the case. A remedy should be considered.

The last element of vector z holds the exact integral I, which is not only dependent on the unknown function but also on the integration domain. The other elements are dependent on the point location and the unknown function, but not on the integration domain. Thus, we decided to manipulate the original system, removing the

last column by subtracting each row to a particular one and holding in vector z only partial derivatives of the unknown function. The particular row selected was the one calculated without removing any data point. As a secondary effect, all elements in the right hand side can be bounded with a multiple of the maximum value of the (p+1)-th derivative (in the 1D problem) or the maximum value of any partial derivative of order (p+1) for 2D and beyond. The trick does not automatically balance the elements of vector z, because the partial derivatives might be very different, but in any cases removes the influence of the integration domain.

Despite having a different unknown vector, we confirmed that the modified system still can be ill-posed. As Hansen [10] explains, the Singular Value Decomposition (SVD) is very useful to illustrate such property. Figure 1 show the singular values for the 2D case N=50, p=8, with point locations created at random. The abscissae varies from 1 up to N*P(p). From the figure it is clear that the spectrum of the base matrix can be divided in three parts: a first and a third one, with exponential decrease of the singular values w.r.t. its index (i.e. a linear function in semi-log plot) with a clear gap in between.

Ill-posed systems require specialized methods for finding a solution. Among the available options, we noticed that the Truncated SVD (TSV) produced good results. As [10] described, it has a direct connection with the traditional Tikhonov regularization problem. It requires selecting in advance a number k of singular values to be retained, which are a property of matrix A and can be calculated once. The Tikhonov approach requires an estimate of the λ parameter, which is a function of A, B and vector f (which might vary from run to run in some application, like daily rain estimate). The criteria can be found in the original reference. Finally, it should be stressed that the SVD calculation is not particularly expensive, but if deemed necessary it can also be stored

for later use provided vector f has no missing values. Otherwise the λ parameter needs to be recalculated.

4. The number of data points is large

In our tests for the 2D case we will show results for N up to 50; otherwise, the computation as implemented will take too much time or exceed the available memory. Such value can be suitable for some meteorological applications but there are a number of others which should deal with over a hundred data points. The proposed algorithm is amenable to cope with large scale problems by noticing the following: if the partial derivatives (or an estimate of them) are available at data points, the computation is straightforward and only requires one Delaunay tesselation plus some computation at each cell. Since the derivatives are local properties, not related with the integration domain they can be obtained using information from its neighbourhood. Thus the procedure could be as follows: the data point set is subdivided in clusters of small size, and for each one the corresponding derivatives can be obtained using our procedure. Such task is trivially parallelizable; once completed, the partial derivative estimate will be available and the global error estimate can be easily built.

5. Numerical tests

We tested the procedure using known analytical functions in 1D and 2D settings. The most interesting case is for higher dimensional problems (2D at least). In practice, the data locations are not completely arbitrary. For example, in meteorological applications data is sparse but its placement has been selected and tested carefully, using more or less sophisticated analysis. The experience shows whether mean areal values obtained with data from the available weather stations are representative of the whole region or not. In addition, the function under consideration can not be completely arbitrary. If it has strong variations for short distances (i.e. low autocorrelation) there is little chance

that sparse data will succeed in accurately estimating the integral. Thus, an experiment where the function is arbitrary and/or the data points are located at random locations might be too pessimistic and not reflect real case performance. So, to be fair, we should test the procedure using functions with reasonable autocorrelation in space, and locations which are realistic.

Data

We selected some toy examples for the case of 1D and 2D. For the former, simple functions like sin(x) and log(1+x) have been considered. The exact integrals can be easily calculated, and compared against the numerical results. For the 2D case, we have selected examples of analytical functions deemed to be representative of some real datasets.

Results for 1D test problem

In Tables 1 and 2 we summarize the results for the 1D case, with equispaced and random locations and N equal to 10, 20, 30 and 40. The chosen integration interval was [5,10]; the value of K was set to 3, and p was selected as 10. The exact solution (I_E) was compared against both the traditional Thiessen estimate (I_T) and our estimate (I_G). In practical settings it is desirable to have a reasonable error estimate: the error is usually computed as the difference of two estimates with different accuracy, like $\Delta I = \left| I_G - I_T \right|$. If we do so, we can confirm experimentally that $I_E \in \left[I_G - \Delta I, I_G + \Delta I \right], \Delta I = \left| I_G - I_T \right|$ in most of the cases and occasionally they are of the same order of magnitude. Table 1 reports the case of regular sampling, while Table 2 evaluates just one example of a random location of data points.

Table 1 Results using equally distributed data points. In bold those cases where the exact value was outside the recommended interval $I_E \in [I_G - \Delta I, I_G + \Delta I]$

function\N	10		20		30		40	
	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E
sin(x)	**1.18e-2**	**5.05e-2**	**2.93e-3**	**2.07e-2**	**1.30e-3**	**4.82e-3**	**7.31e-4**	**2.33e-3**
log(1+x)	7.89e-4	1.81e-4	1.97e-4	1.79e-4	8.77e-5	1.80e-5	4.93e-5	8.02e-6
$(x+1)^2$	0.00e+0	3.55e-14	0.00e+0	2.13e-14	7.11e-15	2.20e-13	0.00e+0	4.12e-13

Notice that the residuals in the last case should be exactly zero, a fact which is approximately attained in all cases once considered the numerical rounding.

Table 2 Results using randomly distributed data points

function\N	10		20		30		40	
	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E
sin(x)	1.24e-2	4.27e-3	2.41e-2	1.04e-2	3.80e-2	5.27e-3	3.33e-3	8.44e-3
log(1+x)	5.78e-3	9.64e-5	7.11e-3	2.86e-4	5.93e-3	1.02e-4	7.90e-4	6.05e-5
$(x+1)^2$	3.84e-2	4.97e-14	5.26e-2	2.84e-14	5.26e-2	1.99e-13	2.63e-4	3.84e-13

In all cases the estimate is closer to the exact value, and in some cases for more than one order of magnitude. The function $(x+1)^2$ is interesting because it has a finite Taylor expansion. This is recognized by the routine and the estimate is fairly closer to the exact value. Figure 2 shows the location of random points, illustrated for the case of fx=log(1+x).

Results for 2D test problem

Table 3 shows the results for the selected functions, which were A= $\sin(x) * \sin(3y)$;

B=$\left(.25-\left(x-.5\right)^2\right)*\sin(\pi y)$ and C=$\left(.25-\left(x-.5\right)^2\right)*y$. The domain was [5,10]x[5,10];

K=6 and p=8. Figure 3 shows the random location of data points.

As before, the error estimate works fairly well, being strict in all but one of the cases, where its magnitude was correctly anticipated.

Table 3 Results for the 2D case using randomly distributed data points. In bold those cases where the exact value was outside the recommended interval

$I_E \in \left[I_G - \Delta I, I_G + \Delta I\right]$

	10		20		30		40		50	
	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E	I_T-I_E	I_G-I_E
A	2.94e-1	2.11e-1	2.96e-1	1.21e-2	1.62e-1	1.15e-3	2.21e-1	2.10e-3	1.56e-1	9.42e-6
B	9.82e+1	2.43e+3	3.30e+2	4.32e+3	2.73e+2	3.18e+2	1.80e+2	7.67e+3	**1.30e+2**	**1.20e+2**
C	6.71e+2	1.58e+2	1.31e+1	4.52e+0	4.61e+1	3.55e+0	6.02e+1	5.16e+0	5.55e+1	8.06e-2

6. Discussion and Conclusions

Areal average of variables of geographic interest has been routinely estimated using the Thiessen's method for more than one century. Despite its popularity, to the best of our knowledge no mathematically sound estimate of its accuracy has been produced so far. Theory and methods coming from numerical analysis are of little use, because they are derived under the assumption that the locations where the function value is known can be prescribed in advance. We devised a new method, which can supplement the traditional Thiessen's estimate with an error bound, valid for functions satisfying reasonable conditions. If the function itself is a polynomial of order up to p the quadrature will be exact and the residual term R_p will be zero disregarding of the data points selected. Thus, we can claim that the new method is *of order p* following common practice in numerical analysis, but keeping in mind that the consequences of

such statement are different. In a traditional setting, having methods of order p and q, p<q is useful in order to estimate an error bound. The absolute value of the difference of quadrature obtained with both methods is used as an estimate of its error, a fact which is valid because the residual is of higher order and it diminishes faster with interval length h. Such conclusion is not fully valid when we cannot diminish h at will.

Despite Thiessen's method has been proposed for two dimensional problems, the procedure can be extended to higher dimensional domains because the rationale is very similar. Our proposal shares such property; we illustrated results for the 1D and 2D case. The MATLAB/OCTAVE code for the 1D and 2D case is available from <send me an e-mail!>

7. Acknowledgements

There are no funding sources to acknowledge, neither conflict of interest to report.

8. References

[1] Thiessen, A. H.: Precipitation Averages for Large Areas. Monthly Weather Review 39, 7, 1082-1084 (1911)

[2] Dahlquist, G. and Björck, Å.: Numerical Methods in Scientific Computing, Vol I, SIAM, ISBN 0898716446, 9780898716443, 746 pp. (2008)

[3] Secrest, D.: Numerical Integration of Arbitrarily Spaced Data and Estimation of Errors. Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis, 2, 1, 52-68 (1965)

[4] Gill, P.E. and Miller, G.F.: An algorithm for the integration of unequally spaced data. The Computer Journal, 5, 80–83 (1972)

[5] NAG: NAG Library Manual Mark 26. The Numerical Algorithms Group, ISBN 978-1-85206-216-3, 12844 pp. (2016)

[6] Singh, A. K. and Thorpe, G. R.: Simpson's 1/3-rule of Integration for Unequal divisions of Integration Domain. Journal of Concrete and Applicable Mathematics, 1, 3, 247-252 (2003)

[7] Minaoui, K., Chonavel, T., Nsiri, B., & Aboutajdine, D.: Quadrature formula for sampled functions. International Journal of Signal Processing, 6, 2, 56-62 (2010)

[8] Efron, B.: Bootstrap Methods: another look at the Jacknife. The Annals of Statistics, 7, 1, 1-26 (1979)

[9] Neumaier, A.: Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization. SIAM Rev., 40, 3, 636–666 (1998)

[10] Hansen, P. C.: Truncated Singular Value Decomposition Solutions to Discrete Ill-Posed Problems with Ill-Determined Numerical Rank. SIAM J. Sci. and Stat. Comput., 11, 3, 503–518 (1990)

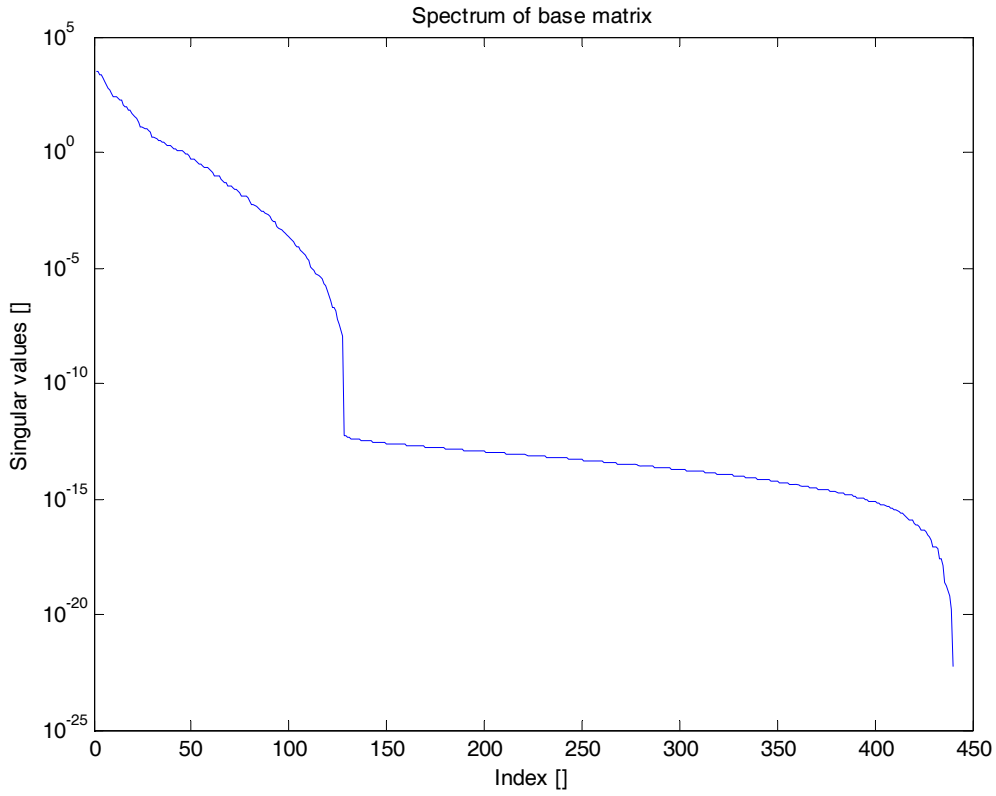

Figure 1 Singular values of the system matrix for the case 2D, N=50 with random coordinates. The gap is located at k=127.

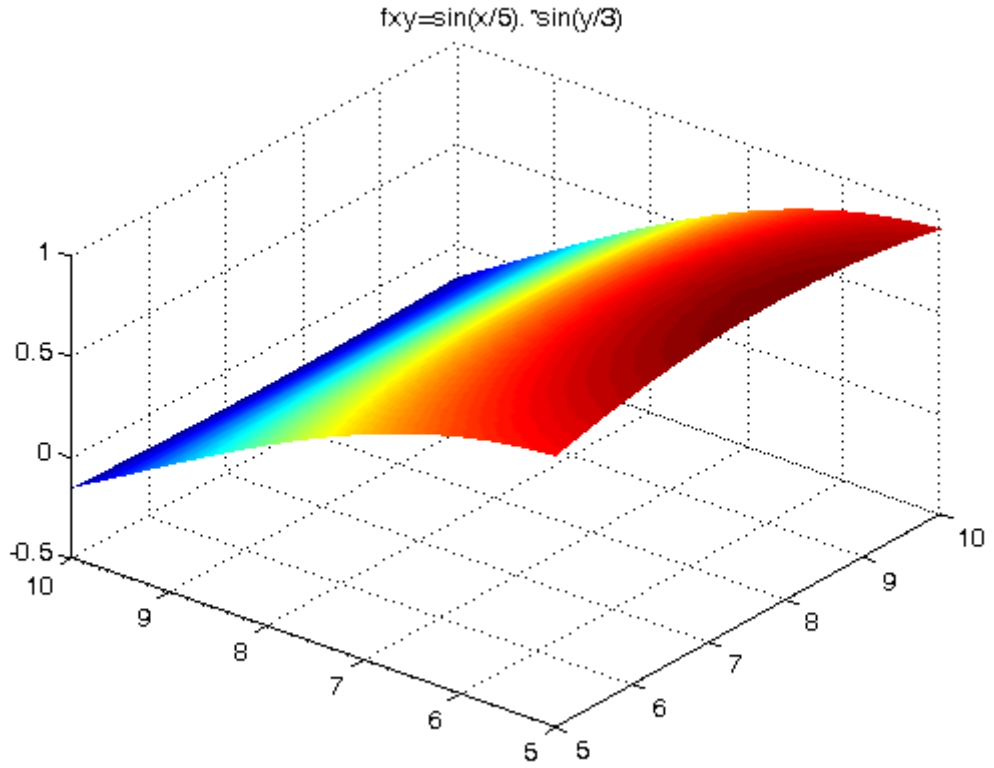

Figure 4 Illustration of the function corresponding to Case A

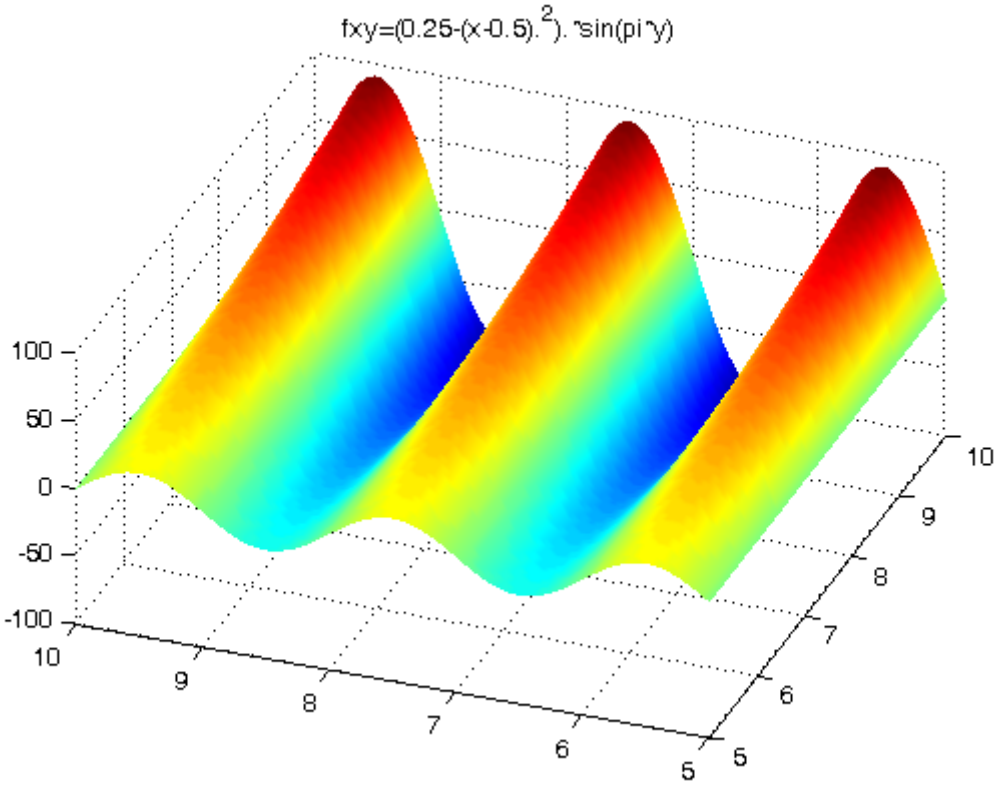

Figure 5 Illustration of the function corresponding to Case B

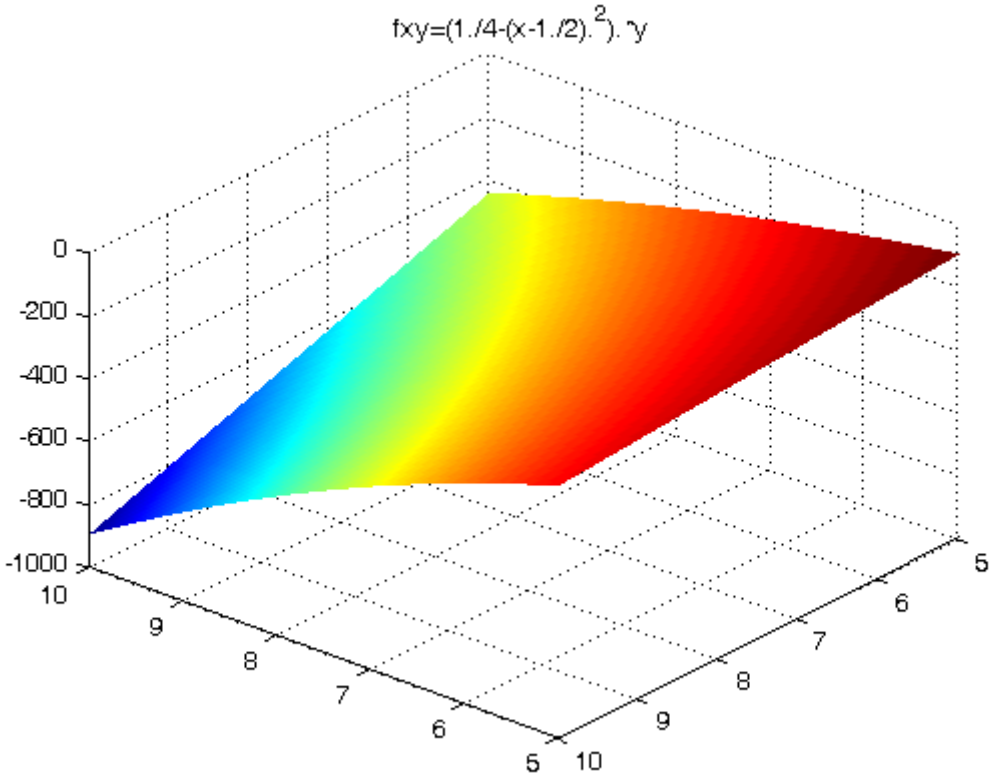

Figure 6 Illustration of the function corresponding to Case C

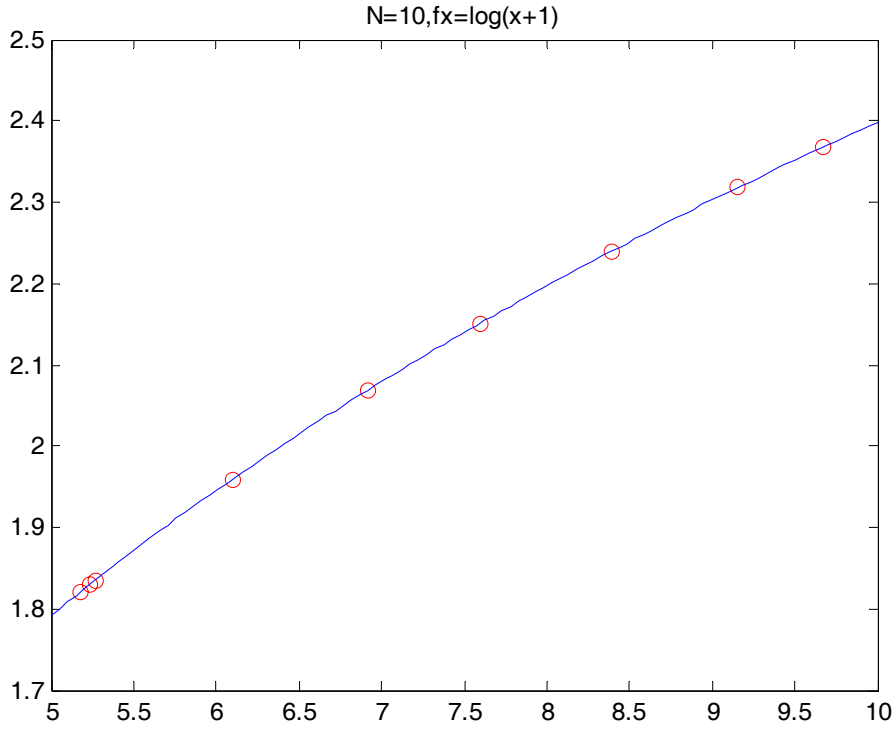

Figure 2 Distribution of 1D data points for the case N=10

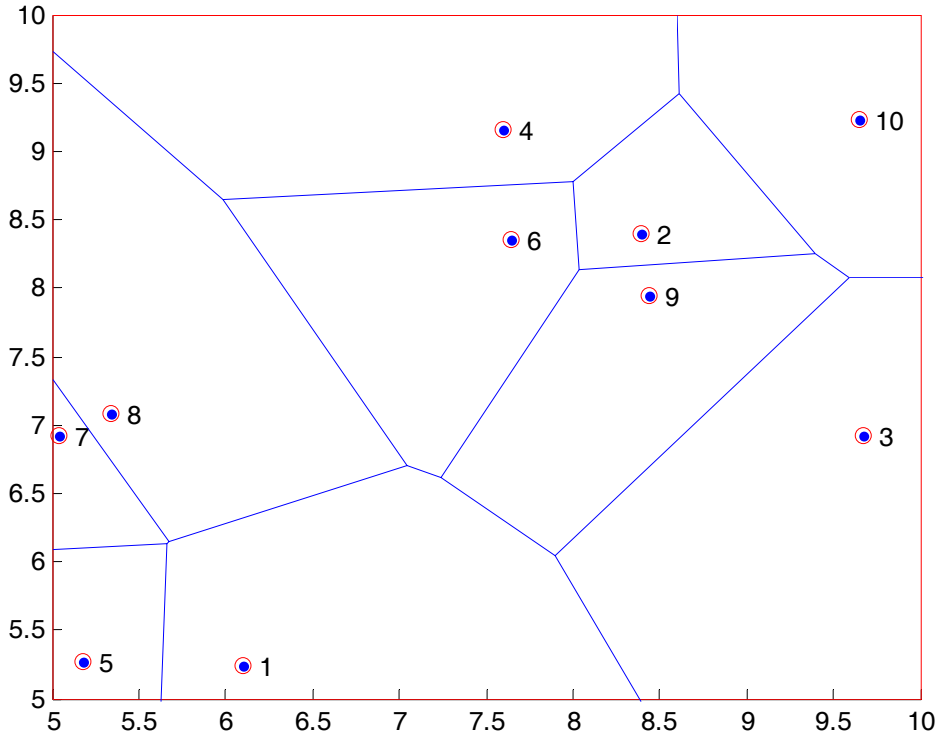

Figure 3 Distribution of 2D data points for the case N=10 and the corresponding Delaunay regions.